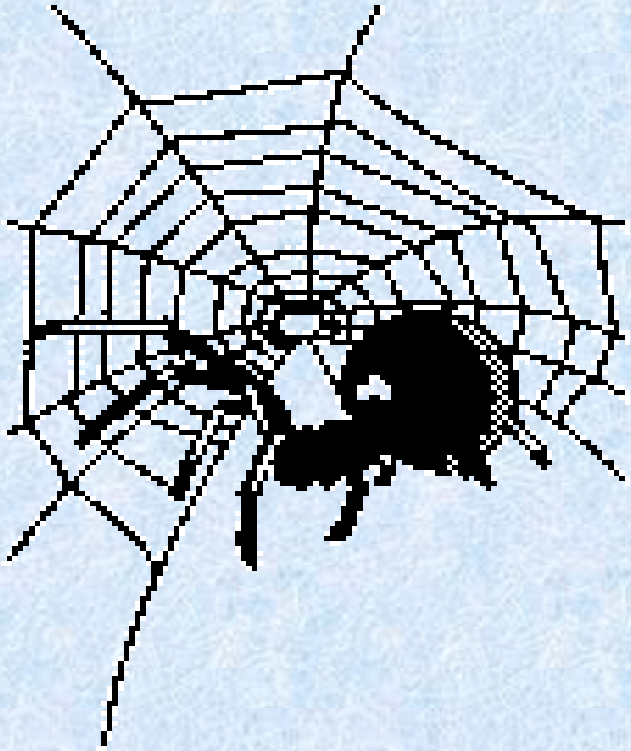


Secrets From the Deep: Mining Government Information from the Invisible Web

Greta E. Marlatt

Dudley Knox Library
Naval Postgraduate School

gmarlatt@nps.navy.mil

SLA/MLD CE
June 8, 2002

Thanks to our sponsors!

http://www.dialog.com/

http://www.ieee.org/

http://www.sla.org/division/dmil/

Field experience is something you don't get until just after you need it.

If it's stupid but it works,

it isn't stupid.

Murphy's Laws of Combat

http://www.swcp.com/~russo/laws-o-combat.html

http://www.jdrcables.com/news/online9.asp#namco

Overview

- Introduction
- What is the Invisible Web?
- Why is it invisible?
- How does one find this information?
- Why do we care about it?

WAL★MART

NEVADA BOYS GOLF

The Invisible Web Defined

- Difficult to define
- Also referred to as "deep web" or "hidden web"

- Content rich databases from universities, libraries, organizations, businesses and the government

- 4 types of invisibility

Opaque web

- depth of crawl
  - expensive to index every page
- frequency of crawl
  - most powerful crawlers can hit only 10 million pages a day
- maximum # of viewable results
- disconnected URLs
  - spiders can’t find

Private Web

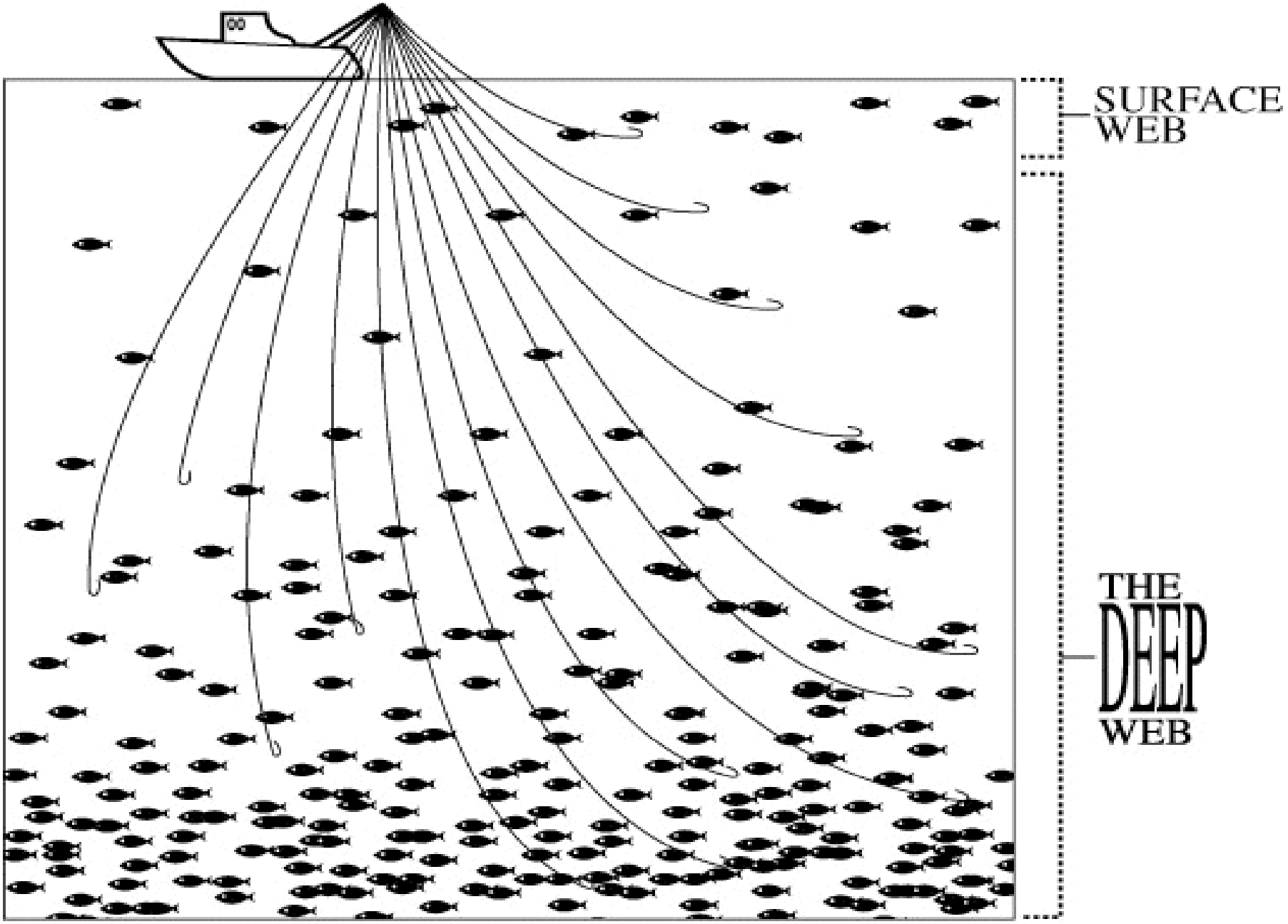
- password protected
- robot.txt file to disallow
- “noindex” meta tag prevents spider from reading past the head portion of the page

Proprietary Web

- registration required
  - some for free
  - some for fee

Truly Invisible

- technical reasons why crawlers can’t find/enter
- search engines have chosen to omit
- dynamically generated pages
- relational databases that require a query



http://home.lausitz.net/a.krause/tiefe_tiefen.htm

What's in the Invisible Web?

- Contents of PDF and other file types [e.g .doc, .xls, .ppt, Flash, streaming media etc.]
  Two sites now index Adobe PDF files:
  http://www.google.com
  http://searchpdf.adobe.com/

- The content of sites requiring registration or a login

- Dynamically generated pages [e.g. CGI, ASP, CFM] where data is requested by a form – also some pages with "?"

Quick Search
Scientific and Technical Documents

This Quick Search searches all databases checked. Choose this search if you want broad-based search results.
The maximum number of documents returned is 25.

Choose one or more of the following :

- [x] DTIC's Technical Reports Collection
- [x] DTIC's Full-Text Collection
- [ ] DTIC Thesaurus
- [ ] R&D Descriptive Summaries (RDDS)

Note: The more you select, the slower the response time.

Search for : View Database Descriptions

"homeland security" Search Clear Query

Fielded Search Search Help

Spiders can't type

Hint : *Secure STINET gives you access to unlimited and limited distribution citations received by DTIC since 1985 as well as full text reports and Research Summaries (RS) reports from 1965 to present. Find out how to* ***register with DTIC and subscribe****.*

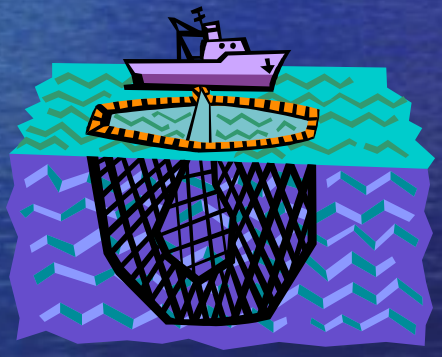
What's in the Invisible Web?

- ✓ Information that resides on an Intranet
- ✓ Commercial resources with domain or IP limitations
- ✓ Sites using a robot.txt file to keep search engines out
- ✓ Real-time data [stock quotes, weather, sports, election results etc.]
- ✓ Graphics [except through the ALT tag]
- ✓ Archives [newspapers]

"stuff" that spiders can't search or won't add

How Big Is It?

- Approximately 500 times larger than the surface web and faster growing
  [BrightPlanet Study 2000]



- 2-50 times larger than the visible web
  [Chris Sherman and Gary Price, The Invisible Web, 2001]

Why Use the Invisible Web?

- Quality of content/higher level of authority
- Comprehensiveness
- Focused
- Timeliness
- The material isn't available elsewhere on the Web

When to Use the Invisible Web

- Standard search engines aren’t working
- A precise answer is needed
- Data or statistics are needed
- High quality or authoritative results are needed

When to Use the Invisible Web

- When timeliness is important
- You know the subject area well
- Looking for collections [images, sounds, manuscripts etc.]
- Reference books online [handbooks, guides, dictionaries, encyclopedias etc.]

Strategies for Searching

- Have the mindset of a hunter, detective or passionate collector

- Use search engines to get to the front door
  - “searchable database”
  - “interactive database”

Strategies for Searching

- Use the site map to see if databases or statistics are mentioned
- Use the site's internal search tool and search for "database"
- Check subject specific discussion groups

Tools

About.com http://websearch.about.com/

Invisible Web Databases -- http://websearch.about.com/cs/invisiblewebdata/

Invisible Web Gateways -- http://websearch.about.com/cs/invisibleweb1/

CALL's Military Domain Search Engine

http://call-search.leavenworth.army.mil/

Searches Multiple (Military related) Sites, including CALL, Army, Army field Manuals, Army regulations, Corps of Engineers, Early Bird, Joint Service, Mitre Corp., NATO, Navy, Rand, USAF, USCG and USMC.

Welcome to CALL'S

Military Domain Search Engine

To narrow your search, specify either text or time peri

Type text to search for:

officer retention

Search for content modified during the last:

choose to search one or multiple catalogs

Use mouse to select one desired catalogs. Ctrl ke mouse selects multiple

- CALL
- CALL-Thesaurus
- Army
- Army_Field_Manuals
- Corps_of_Engineers
- Early-Bird
- Joint-Services
- Mitre
- NATO
- Navy
- Rand
- USAF

Military Domains Search Engine

Approximately 2.3 million records indexed
Click HERE to see the number of records contained in each catalog.

Tips for Searching

For a better understanding of search engines, please read the following article
Effectively Using Search Engines

To use this search engine, you may search all catalogs by simply typing your query, and clicking on search. This searches all available catalogs and time periods. You can further define your search by selecting what catalogs you wish to search and define desired time period when documents were last modified. See below for an explanation

ption to enter a single word or phrase for your search criteria.
ot case sensitive, and it considers both upper and lower case as the

his box blank, provided you specify a time period in the "Content Las
Modified" box.

Content Last Modified

- If left blank, this field will not be considered.
- The drop down menu lets you pick between 24 hours, 7 days, 30 days or 365 days.
- By picking one of those time periods, you limit your responses to documents that were last modified within a specified time frame.
- NOTE: The quality of this field is dependent on the system in which the document resides. If the date was improperly recorded, it will be improperly reported in this area

Catalogs

http://call-search.leavenworth.army.mil/

Tools

CompletePlanet -- http://www.completeplanet.com/

Covers over 103,000 searchable databases and specialty Search engines

Government -- http://www.completeplanet.com/Government/index.asp

Military -- http://www.completeplanet.com/Military/index.asp

http://www.completeplanet.com/

Tools

Direct Search [Gary Price]

http://gwis2.circ.gwu.edu/~gprice/direct.htm

Thousands of specialized databases [link changing]

Price's List of Picks

http://www.freepint.com/gary/listof.htm

Updated: 10/15/2001

direct search

Compiled by:
Gary Price, MLIS
George Washington University
Washington D.C. and Ashburn VA
gprice@gwu.edu

direct search is a growing compilation of links to the search interfaces of resources that contain data not easily or entirely searchable/accessible from general search tools like Alta Vista, Google, or Hotbot. Although these "general" tools are essential for the retrieval of Internet based data, searchers often fail to realize that a massive amount of information is not easily or entirely searchable/accessible via these search tools. Material "hidden" from the general search tools is said to reside on the Invisible Web.

Learn more about the Invisible Web.
Learn more about the OPAQUE/ALMOST INVISIBLE WEB

It's Here!!! A New Book: *The Invisible Web*

Comments and suggestions are welcome.

direct search SearchCenter

This interface provides search access to ALL direct search pages
as well as the following web reference compilations:
Fast Facts
Price's List of Lists

http://gwis2.circ.gwu.edu/~gprice/direct.htm

Tools

FastBoot -- http://www.fastboot.com/mondo_search.html

This site searches over 1,000 search resources broken down into various categories including Politics/Government. Some of the links are to Invisible Web search sites.

Hidden Webs, Specialty Webs and Invisible Web [From FastBoot] -- http://www.fastboot.com/hidden_web.html

First Gov -- http://firstgov.gov/

This is the official gateway to all US Government information searching over 51 million web federal and state web pages.

US government telephone directories and email directories -- http://firstgov.gov/Contact/Directories.shtml

FirstGov Home | About Us | Help | Site Map | June 2, 2002

Search Federal/State

Federal State Both

Search a State

Go

Agencies
- Federal
- State, Local & Tribal
- International

Reference
- News Releases
- Federal Forms
- Laws & Regulations
- Phone Directories
- Questions About Government?
- More

Select a Topic Go

America Responds to Terrorism

Comment to Government

The New FirstGov.gov

Welcome from President Bush

The Official Government Gateways for:

Citizens
interacting with Government

Government helping citizens connect to services and more.

Online Services for Citizens
- Find Government Benefits
- Compare Nursing Homes
- Change Your Address
- Volunteer
- Shopping & Auctions
- Find a Government Job
- Social Security Online
- Apply for Student Loans
- Find a School or Library
- Zip Code Look-Up
- Passport Applications
- Recreation One-Stop
- Birth & Marriage Certificates

Business
interacting with Government

Helping businesses from start to finish with tools and more.

Online Services for Business
- Business Opportunities
- Business Laws & Regs
- Federal Auctions & Sales
- E-File Your Taxes
- Employer ID Number
- Wage Reporting
- Small Business Procurement Registry
- Subcontracting Opportunities
- File Patents & Trademarks
- and much more

E-File Your Taxes

Governments
interacting with Government

Providing governments and their employees with services and more.

Online Services for Governments
- Grants
- Geographic Information
- Employee Locator
- Per Diem Rates
- Federal Personnel-Payroll Changes
- Federal Thrift Savings Changes
- Government Jobs
- Buying Supplies & Services
- Auctions & Sales
- and much more

http://firstgov.gov/

Tools

FirstGov Military -- http://call.army.mil/search/firstgov.htm

Searches the FirstGov web site for items that appears in publicly available military domain web sites.

Flipper.com -- http://www.flipper.com/

Provides true 'Deep Web' search capabilities and categorized results.

GI Search -- http://www.gisearch.com/

For military searches

Google UncleSam -- http://www.google.com/unclesam

Searches only .gov and .mil sites

With this form you can search the FirstGov web site for items that appears in publicly available military domain web sites. If you wish to search all government web sites, please click on one of the graphics at the top of the page to go to the FirstGov homepage.

Search Military Domain Websites

All Words must match

in All Service Branches

(search text) (domain)

Return 10 Detailed search results per page.

Submit

Search Help

This page contains one or more hyperlinks to resources external to the Department of Defense (DoD). The appearance of these hyperlinks does not constitute endorsement by the Center for Army Lessons Learned of these web sites or the information, products or services contained therein. For other than authorized activities such as military exchanges and Morale, Welfare and Recreation sites, the U.S. Army does not exercise any editorial control over the information you may find at these locations. These links are provided consistent with the stated purpose of this DoD web site.

http://
call.army.mil/search/firstgov.htm

Advanced Search
Preferences
Search Tips
ocean waves
Search
http://www.google.com/unclesam
Search the entire web from the Google home page!
Google Home - Advertise with Us - Search Solutions - News and Resources - Language Tools - Jobs, Press, Cool Stuff...
2002 Google
Advanced Search Tips | All About Google
Advanced Search
Find results
with all of the words
database ocean waves
10 results
Google Search
with the exact phrase
with at least one of the words
without the words
Language
Return pages written in
any language
File Format
Only
return results of the file format
any format
Date
Return web pages updated in the
anytime
Occurrences
Return results where my terms occur
anywhere in the page
Domain
Only
return results from the site or domain
.mil
e.g. google.com, .org
More info
SafeSearch
No filtering
Filter using SafeSearch

Tools

GPO ACCESS – http://www.access.gpo.gov/su_docs/index.html

Multi-Database Search -- http://www.access.gpo.gov/su_docs/multidb.html

Description of the databases searched above --

http://www.access.gpo.gov/su_docs/aces/dcff001.html

GovSearch search engine [GlobalMaster] -- http://www.globemaster.de/govsearch/index.html

[.mil & .gov only - combines SearchMil.com and SearchGov.com]

Multi-Database Search

About the Databases
Access Methods
Depository Libraries
FDLP Gateways
Specialized Search
Training Manual

Select one OR MORE databases and enter your search below. There are several ways to search across multiple databases from within a scrolling text box. To perform a fielded search of an individual database, go to the specialized search page. Helpful Hints provide instructions for searching the databases. The default database is the current volume of the *Federal Register*.

- Budget of the US, FY2003
- Budget of the US, FY2002
- Budget of the US, FY2001
- Budget of the US, FY2000
- Budget of the US, FY1999
- Budget of the US, FY1998
- Budget of the US, FY1997
- Campaign Reform Hearing
- Congressional Bills, 103d Congress (1993-1994)
- Congressional Bills, 104th Congress (1995-1996)
- Congressional Bills, 105th Congress (1997-1998)
- Congressional Bills, 106th Congress (1999-2000)
- Congressional Bills, 107th Congress (2001-2002)
- Congressional Directory, 105th Congress (1997-1998)
- Congressional Directory, 106th Congress (1999-2000)

Search Terms:

"embassy bombings"

Maximum Records Returned: 40 Default is 40. Maximum is 200.

SUBMIT CLEAR

Phrases must be in quotation marks (""). Word roots can be searched using an asterisk (*) following the word stem. For example: legislat* will retrieve both legislation and legislative. Additional instructions and examples.

http://www.access.gpo.gov/su_docs/multidb.html

Tools

IncyWincy: The Invisible Web Search Engine

http://www.incywincy.com/

Government -- http://www.incywincy.com/default?p=48669

Military -- http://www.incywincy.com/default?p=371

Infomine Multiple Database Search http://infomine.ucr.edu/search.phtml Infomine might be called an "academic" search engine, focusing on scholarly resource collections, electronic journals and books, online library card catalogs, and directories of researchers.

Government Information –

http://infomine.ucr.edu/cgi-bin/search?govpub

IncyWincy
The Invisible Web Search Engine
+earthquake +monitor
Search
(advanced)
Arts
Movies, Television, Music...
Business
Jobs, Industries, Investing...
Computers
Internet, Software, Hardware...
Games
Video Games, RPGs, Gambling...
Health
Fitness, Medicine, Alternative...
Home
Family, Consumers, Cooking...
Kids and Teens
Arts, School Time, Teen Life...
News
Media, Newspapers, Weather...
Recreation
Travel, Food, Outdoors, Humor...
Reference
Maps, Encyclopedias, Dictionaries...
Regional
US, Canada, UK, Europe...
Science
Biology, Psychology, Physics...
Shopping
Autos, Clothing, Gifts...
Society
People, Religion, Issues...
Sports
Baseball, Soccer, Basketball...
World
Deutsch, Español, Français, Italiano, Japanese, Korean, Nederlands, Polska, Svenska...
Help build the largest human-edited directory on the Web.
Submit a Site - Open Directory Project - Become an Editor
Powered by Net Research Server
http://www.incywincy.com/

http://infomine.ucr.edu/

Tools

InvisibleWeb.com -- http://www.invisibleweb.com/

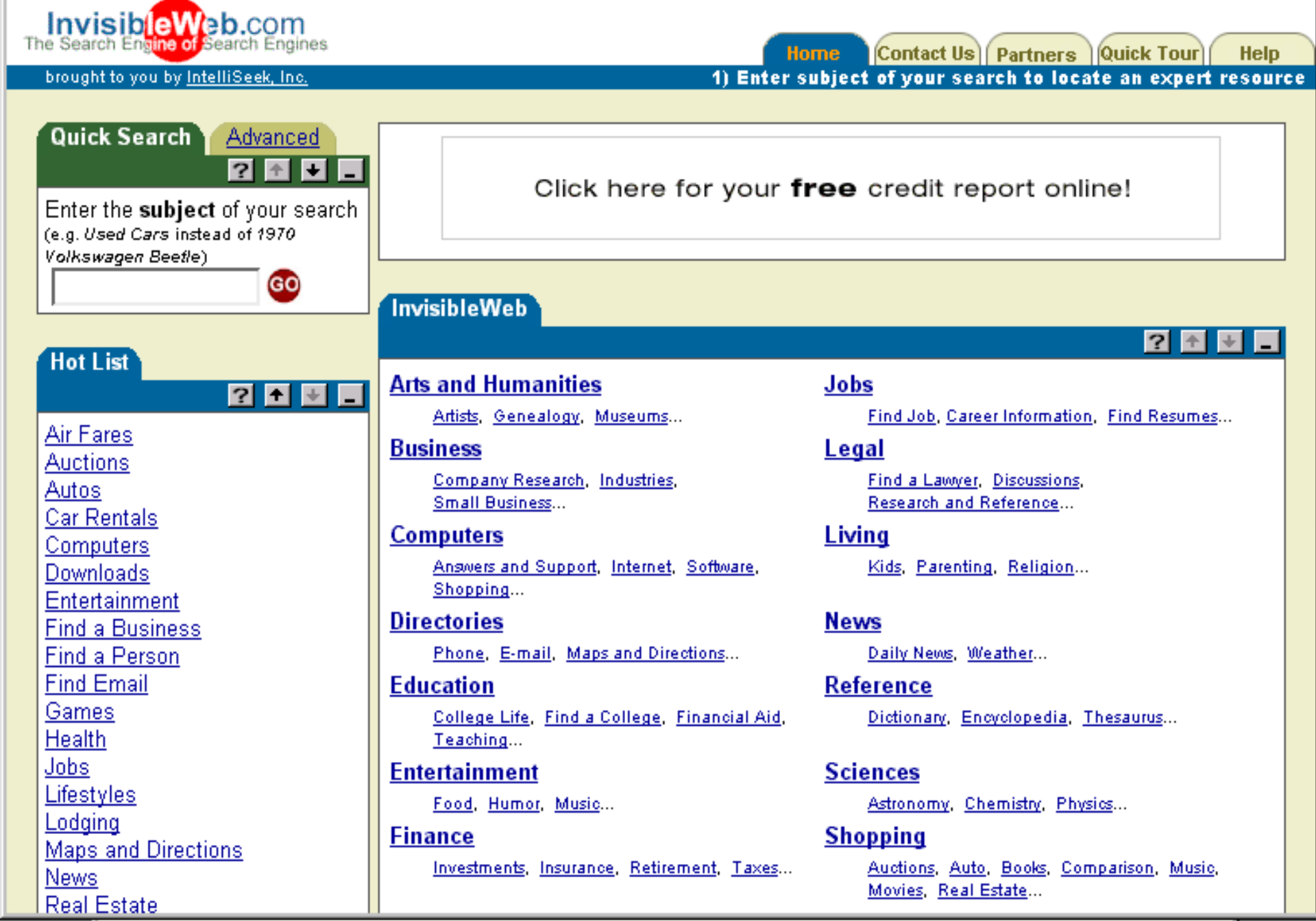
A directory with over 10,000 search engines organized into categories.

Invisible Web Directory -- http://invisible-web.net/

This site is a companion to ***The Invisible Web: Finding Hidden Internet Resources Search Engines Can't See*** by Chris Sherman and Gary Price.

Jane's Defence Discovery -- http://www.defence-discovery.com/

This is a free specialist search engine which searches military government, defense industry and non- governmental organizations (NGO's)



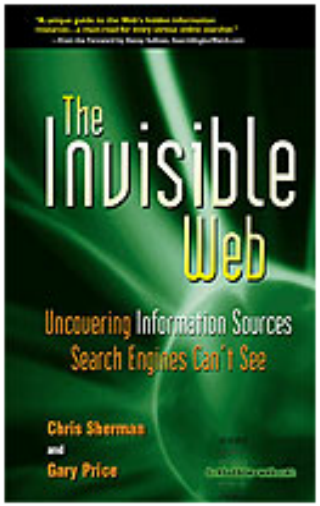
http://www.invisibleweb.com/

www.invisible-web.net

About the Site

About
FAQ
Suggest a Site

About the Book



Overview
Contents pdf
Introduction pdf
Chapter 12
About the Authors
Buy the book

The Invisible Web Directory

Art and Architecture
Bibs/Library Catalogs
Business and Investing
Computers and Internet
Education
Entertainment
Government Info
Health and Medical
Legal and Criminal
News and Current Events
Public Records
Real-Time Information
Reference
Science
Searching for People
Social Sciences
Transportation
U.S./World History

Tools

Librarian's Index to the Internet [LII] -- http://lii.org/

Government -- http://lii.org/search/file/government

Military -- http://lii.org/search/file/military

Navigating the Military Internet http://web.nps.navy.mil/~library/milinternet.htm

This page, while not a search tool per se, has a lot of good links to all kinds of military information. A related page Where to Find Military Information is also useful since it is organized by topic and leads to many databases containing military information. -- http://web.nps.navy.mil/~library/militaryinfo.htm

Librarians' Index to the Internet

lii.org

Information You Can Trust

- About the LII
- Browse *All* Subjects
- Best Search Tools

patents | Search | Help | Advanced Search

Search by: (•) All Fields ( ) Subject ( ) Title ( ) Description ( ) URL [ ] No Stemming

Results for patents 1 to 20 of 26 (view all) Top 20 subjects

View: Next 6

Patent Databases (US Patent and Trademark Office) score: 1000

Search the Patents databases (full-text with images or just bibliographic with quick information) from 1976 to most recent weekly issue. Returns bibliographic data, abstract, and full text of the patent with drawings or photos. Patents from 1790-1976 are also availble, but are searchable only by patent number and current US classifications. Images require ability to view G4 TIFF format. The site provides information on a free version of such a viewer on their Document Formats and Viewers page *(on the Web4Lib mailing list, librarians have recommended the AlternaTIFF viewer for Windows machines)*. For patent offices in other countries see their Related Sites page.

http://www.uspto.gov/patft/

Subject: Patents

Category: Best of...

Created by: cl on May 8, 1999 - updated Jan 4, 2001 | Comment On This Record

Patent Law Links score: 1000

Comprehensive site for all aspects of patent work. Includes links to case law, statutes and regulations, U.S. Patent Office, patent drawing services, patent law firms, patent search firms, document retrieval firms, intellectual property law and technical schools, professional associations and journals, and legal search engines. U.S. Patent Office forms can be read and printed with Adobe Acrobat Reader.

http://www.patentlawlinks.com/

Subject: Patents

Category: Directories

Where to Find Military Information

Dudley Knox Library
Naval Postgraduate School

Ask a Librarian | BOSUN | Databases | Ejournals | Interlibrary Loan | Resources | Web Tools

FT - Full-text

- NPS Domain Only

NPS Modem Pool Access
Proxy Server Access

Awards, Insignia, Medals, Seals, Coins
- Awards, Decorations and Devices
- Badges/Insignia/Patches/Ratings/Wings
- Grade/Rank/Rate Insignia
- Military Aircraft Fin Flashes & Tail Markings
- Records, Trophies and Competitions
- Seals/Logos/Shields/Coins

Budget and Expenditures
- Budget Breakdowns – Various Defense Budgets
- Defense Contractors and Procurement
- Foreign Military Sales and Training
- Pay Schedules -- Civilian [GS & SES]
- Pay Schedules -- Military
- Pay Schedules -- Retired and annuitant pay
- Weapons Procurement Statistics
- Where Military Dollars Are Spent

Education, Research & Medicine
- Military Medicine

Manpower
- Career Fields/Occupational Specialties
- Defense Presence by State
- DoD Civilian Work Force
- Enlisted Totals & Breakdowns
- Force Structure and Manpower Strengths
- Guard & Reserves
- Military Families
- Military Retirees
- Occupational Field/Specialties Population
- Officer Totals & Breakdowns
- Retention/Re-Enlistment Rates
- Where They Serve

Organization
- Field Operating Agencies and Command Locations
- Leadership (Including Chronologies)
- Military Installation Address and General Info
- Military Installation Maps
- Order of Battle

http://
web.nps.navy.mil/~library/militaryinfo.h

Tools

Oingo [a “meaning based” search tool] -- http://www.oingo.com/

Society > Government -- http://www.oingo.com/topic/57/57764.html

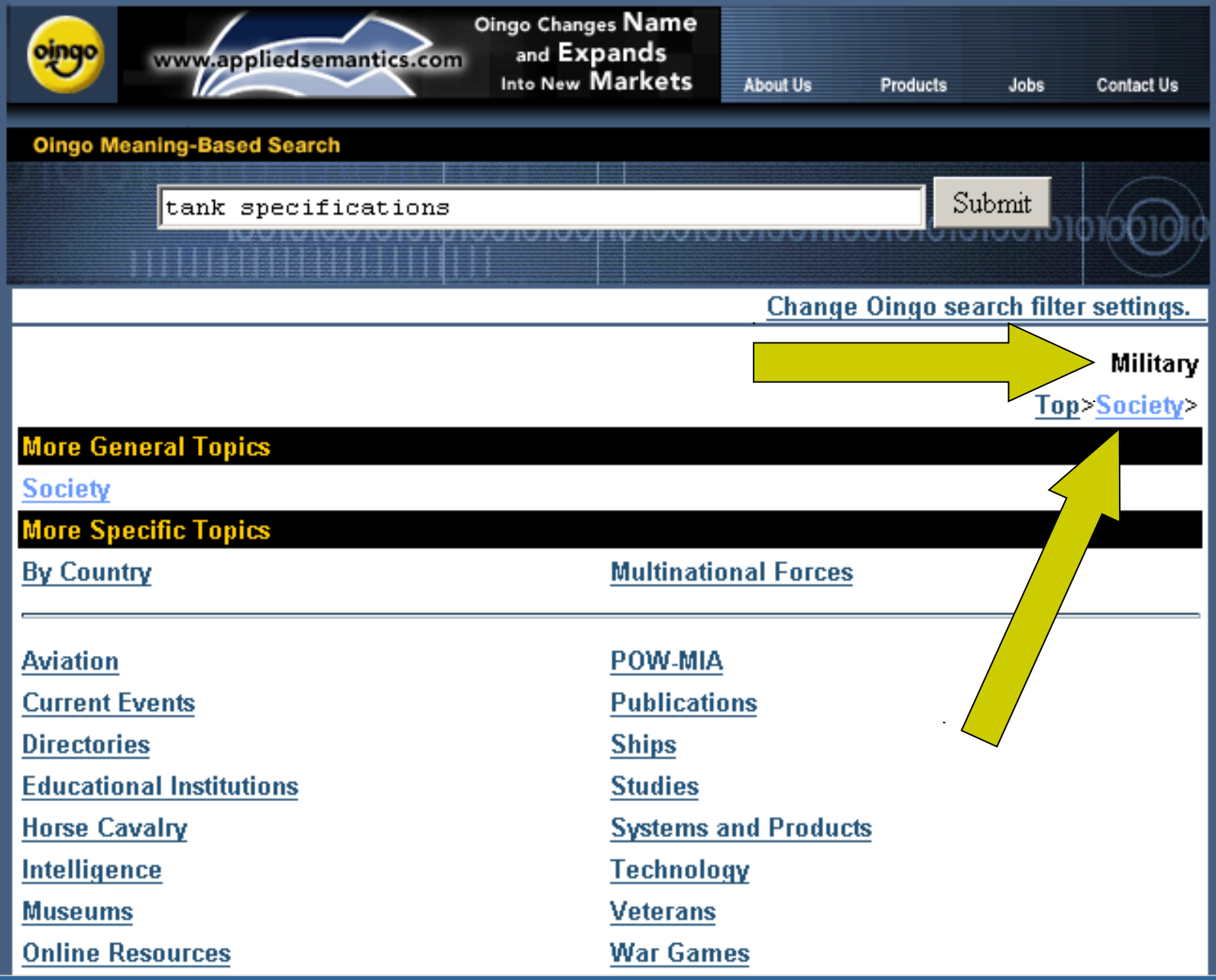
Society > Military -- http://www.oingo.com/topic/53/53534.html

ProFusion -- http://www.profusion.com/

When using their directory to drill down, Invisible web sites are denoted with a + symbol.

Government [including Defense] -- http://www.profusion.com/AgentNav.asp?PTH=ProFusion/Government&TTL=Government

Invisible Web -- http://www.profusion.com/CatNav.asp?PTH=ProFusion/Web&TTL=Invisible%20Web%20Sites



http://www.oingo.com/topic/53/53534.html

sign in

ProFusion
Powered by Intelliseek

home | alerts | options | help | feedback

What's new in ProFusion?

Search the Top 1000 Search Sources and get the BEST results!

Search for: inflation rate [Find It!] Advanced

>> Click here to go to the classic ProFusion advanced web search. <<

Topic Searches

Target your search using these groups of vertical search sources.

Arts & Humanities
Art, Architecture, Archaeology...

Business
Business News, Small Business, Company Profiles...

Career
Find Jobs, Career Tips, High-Tech Jobs, Resumes...

Developer Toolbox
Developer Resources, HTML, ASP...

Entertainment
Movie Reviews, Television, Celebrities...

Finance
Finance News, Ticker Symbol Look[illegible], Mutual Funds...

Government
Federal Government, State & Local Government, Tax Forms...

Health
Health Tips, Drugs, Alternative Medicine, Medical Terms...

News
Headlines, Business News, Tech News, Sports News...

[illegible]nce
[illegible]graphies, Encyclopedias, PDF...

Science
Biology, Agriculture, Astronomy...

Sports
Baseball, Basketball, Football, Hockey...

Take the ProFusion Quick Tour

Make ProFusion your home page!

Search Tip

Want the Power of Profusion for your Company?

Presenting Intelliseek's Enterprise Search Server -- Intelligent Knowledge Discovery Across Enterprises

Click here for more information.

http://www.profusion.com/

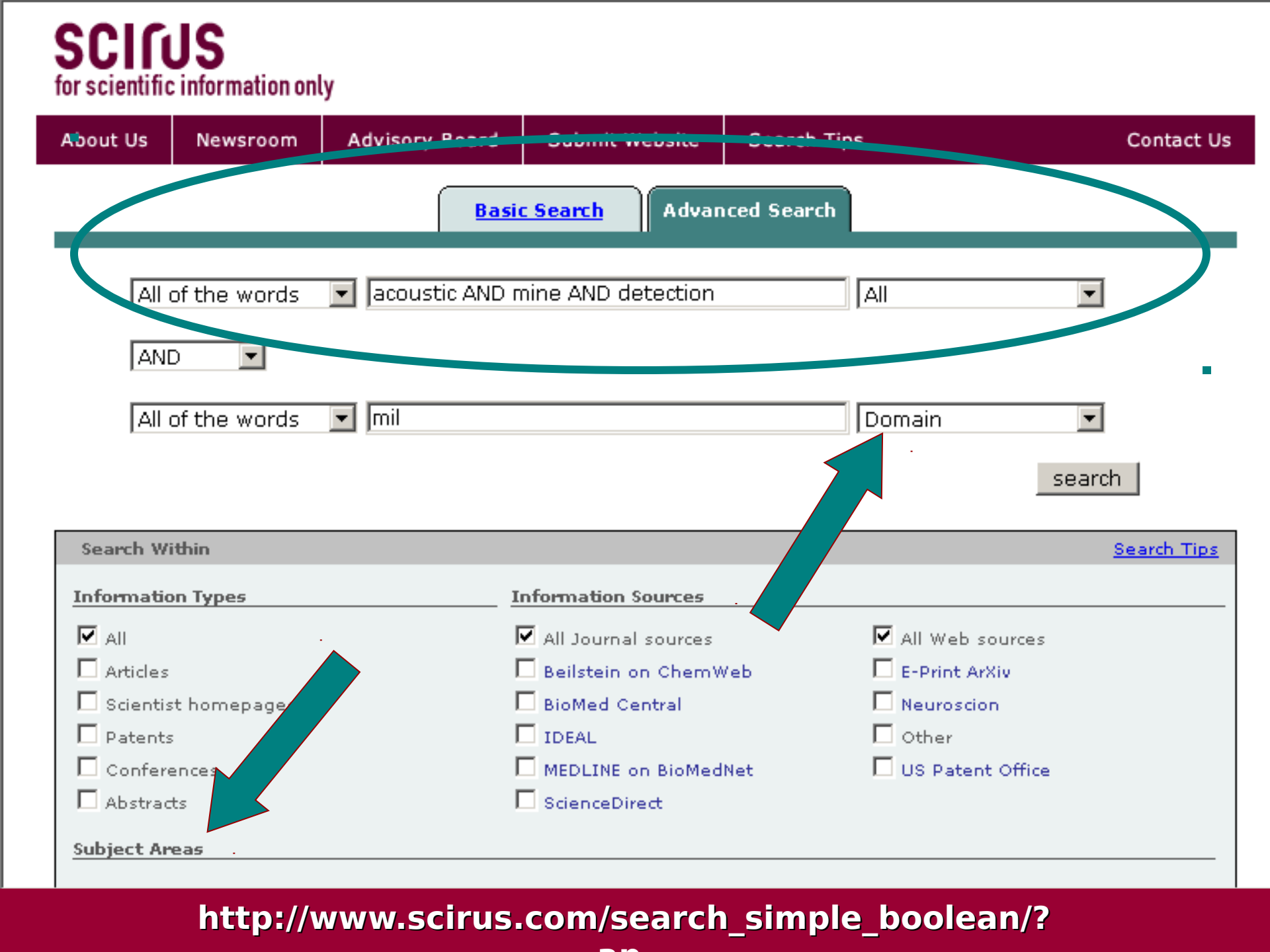
Tools

Scirus -- http://www.scirus.com/?ap

This site concentrates on scientific information. It searches both free and journal sources. Their spiders crawl deeper and do recognize non-text files such as PDF.

SearchGov.com -- http://www.searchgov.com/

SearchMil.com -- http://www.searchmil.com/



http://www.scirus.com/search_simple_boolean/?

Tools

USGOVSEARCH -- http://usgovsearch.northernlight.com/

Search options include Northern Light's Special Collection, NTIS Archive, government-related WWW

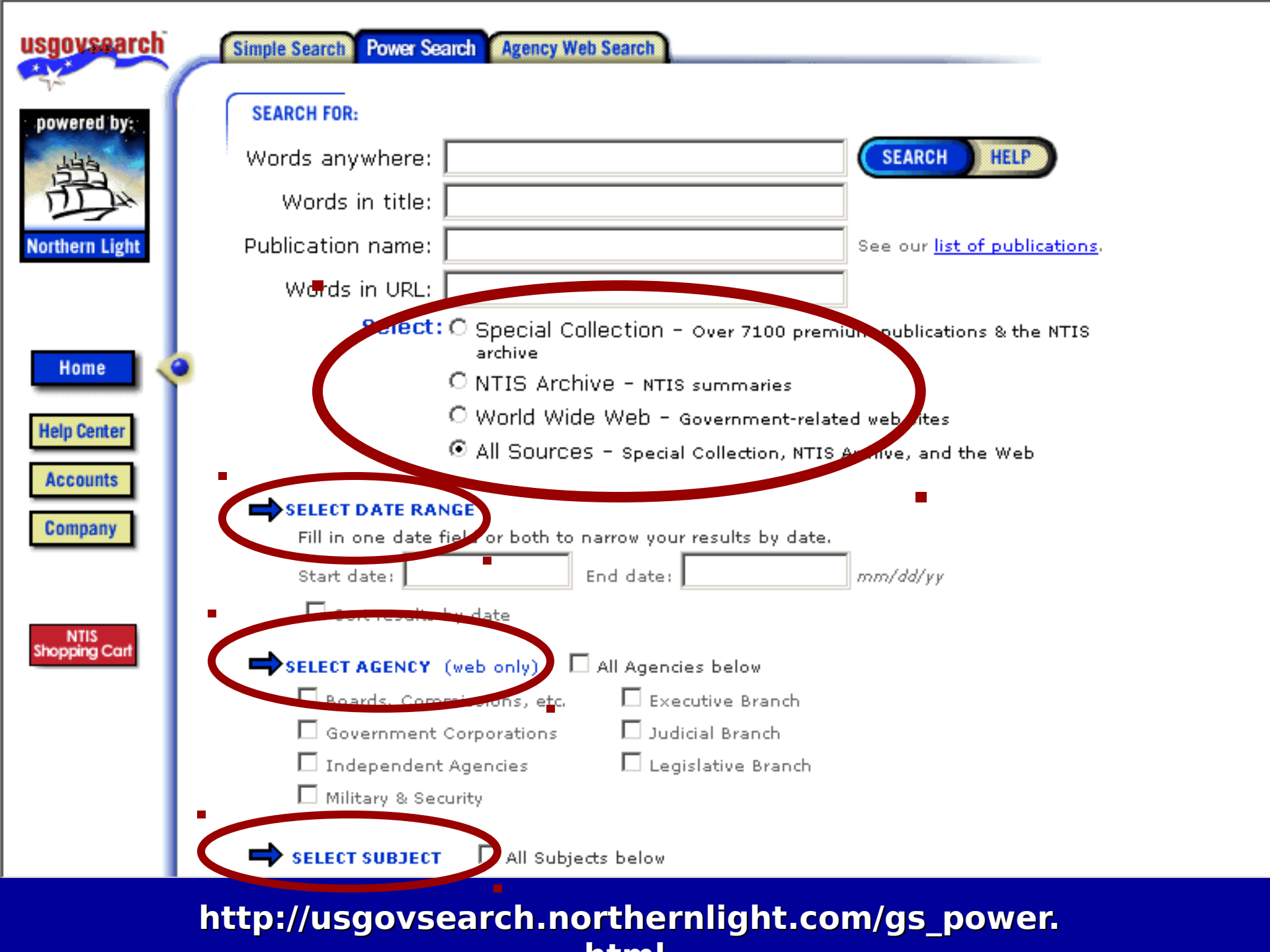
Power search -- http://usgovsearch.northernlight.com/gs_power.html

Agency web search -- http://usgovsearch.northernlight.com/gs_agency.html

WebData.com -- http://www.webdata.com/webdata.htm

Government -- http://www.webdata.com/pnum/Classification7.htf

Military -- http://www.webdata.com/pnum/SubClassification711.htf



http://usgovsearch.northernlight.com/gs_power.html

Staying Informed

- Read publications that review web sites
  - Computers in Libraries
  - Online magazine
  - Searcher
  - Subject specific journals
- Subscribe to discussion groups in your specific areas of interest

Staying Informed

- ResearchBuzz -- Tara Calishain
  http://www.researchbuzz.com/

- Free Pint Newsletter
  http://www.freepint.com

- Virtual Acquisition Shelf & News Desk [aka "Resourceshelf" Gary Price]

http://resourceshelf.freepint.com/

Staying Informed

- Librarian's Index to the Internet
  http://lii.org
- LLRX.com Newsletter http://www.llrx.com/subscribe.htm
- CPL Internet Gazette
  http://www.suffolk.lib.ny.us/libraries/cmmk/gazette_main.html
- Scout Report
  http://scout.cs.wisc.edu/

Staying Informed

- **GPO's New Titles by Topic E-mail Alert Service**

  http://bookstore.gpo.gov/alertservice.html

  This free service allows you to receive convenient and timely e-mail updates of selected new publications available for sale from the Superintendent of Documents. Business, Defense and Security, Elementary and Secondary Education, Employment and Occupations, Federal Statistics, Health Care, Military History, Travel and Tourism.

- **Infominder**

http://www.infominder.com/webminder/index.jsp

Infominder is a free web product that will inform the user of new additions to any web site via e-mail. Infominder provides a link back to the site with the additions highlighted.

Questions?

CREDITS

Welcome Spider
http://library.rider.edu/scholarly/rlackie/Invisible/Inv_Web.html

Chief of the Watch -- U571 movie sound file
http://www.thewavplace.com/misc/movies6/u571_dalarm.wav

Dive Alarm [Goatlocker Org.]
http://www.goatlocker.org/wavfiles.htm

Sonar Ping [Goatlocker Org.]
http://www.goatlocker.org/wavfiles.htm

Surface, Surface [Goatlocker Org.]
http://www.goatlocker.org/wavfiles.htm

Roaming Spiders -- Animation Arthouse: Spiders